Meta-analysis on continuous outcomes in minimal important difference units

an application with appropriate variance calculations

Shrier, Ian; Christensen, Robin; Juhl, Carsten Bogh; Beyene, Joseph

Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2016.07.012

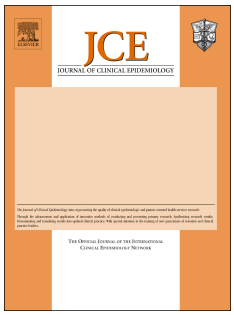
Publication date:
2016

Document version
Peer reviewed version

Document license
CC BY-NC-ND

Citation for pulished version (APA):
Shrier, I., Christensen, R., Juhl, C., & Beyene, J. (2016). Meta-analysis on continuous outcomes in minimal important difference units: an application with appropriate variance calculations. Journal of Clinical Epidemiology, 80, 57–67. DOI: 10.1016/j.jclinepi.2016.07.012

Accepted Manuscript



Meta-analysis on continuous outcomes in minimal important difference units: An application with appropriate variance calculations

Ian Shrier, MD, PhD, Robin Christensen, MSc, PhD, Carsten Juhl, PT, MPH, PhD, Joseph Beyene, MSc, PhD

PII: S0895-4356(16)30259-1

DOI: 10.1016/j.jclinepi.2016.07.012

Reference: JCE 9217

To appear in: *Journal of Clinical Epidemiology*

Received Date: 10 August 2015

Revised Date: 29 June 2016

Accepted Date: 2 July 2016

Please cite this article as: Shrier I, Christensen R, Juhl C, Beyene J, Meta-analysis on continuous outcomes in minimal important difference units: An application with appropriate variance calculations, *Journal of Clinical Epidemiology* (2016), doi: 10.1016/j.jclinepi.2016.07.012.

Meta-analysis on continuous outcomes in minimal important difference units: An application with appropriate variance calculations

Ian Shrier MD, PhD; Robin Christensen MSc, PhD; Carsten Juhl PT, MPH, PhD; Joseph Beyene MSc, PhD

Affiliations

IS: Centre for Clinical Epidemiology, Lady Davis Institute, Jewish General Hospital, McGill University, Montreal, Canada.

RC: Musculoskeletal Statistics Unit, The Parker Institute, Dept. Rheum.; Bispebjerg and Frederiksberg Hospital, The Capital Region of Copenhagen, Denmark.

CJ: Research Unit for Musculoskeletal Function and Physiotherapy, Department of Sports Science and Clinical Biomechanics, University of Southern Denmark and Department of Orthopedics, Department of Orthopedics, Copenhagen University Hospital, Herlev and Gentofte, Denmark.

JB: Department of Clinical Epidemiology and Biostatistics, and Department of Mathematics and Statistics, McMaster University, Hamilton, Canada

Address Correspondence To:

Ian Shrier MD, PhD

Centre for Clinical Epidemiology

Lady Davis Institute for Medical Research, Jewish General Hospital

3755 Cote Ste-Catherine Road

Montreal, QC H3T 1E2

Tel: 514-340-8222 ext 4244

Fax: 514-340-7564

ian.shrier@mcgill.ca

Word Count: 4690 words

Abstract Word Count: 184

ABSTRACT

Objective: To compare results from meta-analyses for mean differences in minimal important difference (MID) units (MD_{MID}), when MID is treated as a random variable versus a constant.

Study Design and Setting: Meta-analyses of published data. We calculated the variance of MD_{MID} as a random variable using the delta method, and as a constant. We assessed performance under different assumptions. We compare meta-analysis results from data originally used to present the MD_{MID}, and data from osteoarthritis studies using different domain instruments.

Results: Depending on the data set and depending on the values of rho and CoV-MID, estimates of treatment effect and p-values between an approach considering the MID as a constant versus as a random variable may differ appreciably. Using our data sets, we provide examples of the potential magnitude. When rho=0.5 and CoV_{MID}=0.8, considering MID as a constant overestimated the treatment effect by 33%-110%, and decreased the p-value for heterogeneity from above 0.95 to below 0.08. When rho=0.8 and CoV_{MID}=0.5, the magnitude of the effects were similar.

Conclusions: Considering MID as a random variable avoids unrealistic assumptions and provides more appropriate treatment effect estimates.

Keywords: continuous outcomes; meta-analysis; minimal important difference; standardized mean difference; ratio of means; variance; methods

What is new?

Key finding

Current methods to standardize continuous outcomes in minimal important difference units require unrealistic assumptions

What this adds to what was known?

We describe a method to standardize continuous outcomes in minimal important difference units that allows for greater transparency of assumptions, and sensitivity analyses.

What is the implication, what should change now?

When standardizing continuous outcomes using minimal important difference units, investigators should incorporate realistic assumptions.

Investigators should use sensitivity analyses to test the robustness of the results to violations of their assumptions.

Some examples for GRADE Summary of Findings tables are provided.

INTRODUCTION

Health care professionals are strongly encouraged to practice evidence-based medicine (EBM) [1-4], where clinical decisions are based on the best evidence addressing a focused clinical question. Practicing EBM requires access to health care evidence, and preferably evidence that is succinctly and systematically summarised. When there is sufficient homogeneity of studies, a meta-analysis fulfils this objective. A meta-analysis may be broadly defined as the quantitative review and synthesis of the results from related but independent studies [5]; a trustworthy meta-analysis is always based on a thorough systematic review of the literature wherein the authors provide an overall quantitative summary statistic for the effect estimate of a group (or subgroup) of studies [6].

When the outcome is binary, investigators commonly combine studies in a meta-analysis by choosing to summarize across the risk difference, risk ratio or odds ratio scales [7, 8]. The magnitude and direction of the overall estimate may be different with different summary statistics because the formula for the variances (responsible for the weighting of individual studies) are different; the appropriate summary statistic for a particular meta-analysis may depend on the underlying reasons for variation in control group event rates; in some situations uncertainty about the choice of summary statistic will remain [9]. Therefore, to avoid introducing reporting bias, investigators should be explicit about why they chose the particular summary statistic for binary data [10].

When the outcome is continuous, systematic reviewers must calculate the treatment effect as either a raw mean difference (MD), or standardize the mean difference in some way [5]. Standardizing the mean difference is typically preferable when the construct being measured is the same across studies, but the actual measurement instrument differs. For example, frequently used pain measures for osteoarthritis [11] include the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) [12], Knee Injury and Osteoarthritis Outcome Score (KOOS)/ Hip Disability and Osteoarthritis Outcome Score (HOOS) [13, 14], Visual Analogue Scales, Health Assessment Questionnaire (HAQ) (pain subscale) [15], Lequesne algofunctional index (pain subscale) [16], Arthritis Impact Measurement Scales

(AIMS) (pain subscale) [17, 18] and McGill Pain Questionnaire (pain intensity) [19]. When constructs are measured using different scales, combining the raw numbers into a weighted average is not meaningful because a result of 10 on one scale might be equivalent to a result of 50 on another scale. Therefore, some form of standardization is necessary before the results can be combined.

Commonly proposed methods for standardization include the standardized mean difference (SMD) [20], ratio of means (RoM) [21, 22], and a more recent method based on standardizing the MD using minimal important differences (MID) between groups [23, 24]. Although the MID approach has recently been proposed as a simple effect measure to use, it considers MID as a constant. However, different patients will often have different values for MID, just as different people have different heights or weights. For example, if pain is rated on a scale of 0-10, one person might consider 2 as the MID, another 3 as the MID and another 1 as the MID. If we acknowledge that there is variation in the population, then in statistical terms, the MID is considered a random variable; taking the mean of the values as the one true value would be to treat MID as a constant.

The distinction between treating MID as a constant versus a random variable is important. As a random variable, there would be an expected correlation between MD and MID, and there would be a coefficient of variation of MID. The value of these two variables will affect confidence intervals and statistical significance testing. The purpose of this paper is to highlight the overall benefits of treating MID as a random variable, and to illustrate how it can be implemented easily. We illustrate our proposed solution using two different datasets, (i) data originally pooled by Johnston et al. in their original study [23], and (ii) data from studies investigating the effects of exercise on knee osteoarthritis [25] that are well-known for using different scales to measure the same construct [11].

Finally, once the treatment effect is estimated (whether SMD, RoM, or MID units), authors have different options on how to present the results. The GRADE working group have suggested that summary of findings tables could include a comparative treatment effect such as mean difference, mean values for each group (by assuming a mean value for the control group and then estimating the mean for the treatment group based on the calculated treatment effect), or converting the continuous scale into

categories and reporting proportion of patients who would receive substantive benefit [26]. Although the objective of our paper is to estimate a valid treatment effect, the editors have asked us to illustrate how the results could be adapted into different formats for presentation to decision-makers.

Proposed Standardization Methods

The most commonly recommended and used method of standardization is the SMD [20, 27]. In brief, the mean difference from each study is divided by a pooled standard deviation (sd). Therefore, each study estimate is now expressed as an "effect size", and these can be combined using standard inverse-variance meta-analytical techniques. As discussed by the GRADE working group, the problem with this method is that two studies will have different effect estimates if the sd differs [26], even if they use the same scale, have the same mean difference, and have the same number of participants. Therefore, a study with a larger sd not only receives less weight (may sometimes be appropriate but can result in a bias towards the null [27]) but also has a reduced effect size and this leads to bias. Imagine a series of studies where all studies measured an outcome using two different scales (Scale A, Scale B). If one conducted a meta-analysis based on mean difference without standardization using only Scale A or only Scale B, the results are expected to be similar as long as both scales are measuring the same construct accurately. Now, consider that 50% of the studies lost their Scale A data (or never collected Scale A data), and the other 50% of the studies lost their Scale B data (or never collected Scale B data). The meta-analysis based on standardized mean differences would yield different (biased) results from the original meta-analysis using either Scale A or Scale B because the effect estimates are now dependent on the sd of the individual studies. In response to these challenges, two alternatives have recently been proposed.

Friedrich et al. proposed to base the meta-analysis on the ratio of means (RoM) of the two groups [21, 22]. As a ratio, the effect estimate from each study has no units and can be combined. Further, the multiplicative nature of the RoM (e.g. the treated group has ½ the pain of the untreated group) is appealing for clinicians and patients because treatments are often discussed in these terms. However, this

is only appropriate if the treatment effect indeed occurs on the multiplicative scale. For example, a treatment might reduce pain by 4 units on a 10-point scale whether the baseline pain is 8 (final pain score = 4, RoM = 2) or 6 (final pain score = 2, RoM = 3), i.e. mean difference equals 4 for both studies. In this case, an additive scale would be less heterogeneous and more appropriate than the multiplicative scale of RoM.

In 2010, Johnston et al. [23] proposed to standardize the mean difference (additive scale) by the MID. In this measure, the results from each study are represented as the number of MID units for the scale that was used. In the example above with a mean difference of 4, if the MID for the pain scale was 2, both studies would decrease pain by 2 MID units. This standardization is appealing to clinicians for contexts requiring an additive scale because it has face validity and is easily interpretable. However, there are two challenges with Johnston et al.'s approach: 1) the calculation of the variance of the MD/MID ratio, and 2) many studies do not have an established MID.

Obtaining a valid variance of individual studies in a meta-analysis is essential because variance affects both the point estimate of the overall meta-analysis estimate (dependent on a weighted average of the individual studies, where the weights for each study are based on their respective variances), the observed heterogeneity, and the uncertainty of the overall estimate (dependent on the sum of the individual variances in a fixed effects analysis, and the sum of the within and between study variances in a random effects analysis). To obtain the variance of the MID unit effect estimate, Johnston et al. considered the MID for a scale to be a constant [23].

The variance of an estimate (e.g. mean difference) divided by a constant (e.g. MID) is simply the variance of the estimate divided the constant squared (MID^2). Simplifying MID to a constant (i.e. one value) is a common practice, but introduces challenges. First, establishing an MID for a scale is difficult because the MID for one patient is different from another patient. Further, the MID may be dependent on the baseline value prior to treatment [28, 29]. Therefore, treating the MID as a constant appears to be an unrealistic assumption. Second, a meta-analysis using MID units requires that the MID be defined for each scale in the meta-analysis. Unfortunately, the MID for many scales has not been formally

investigated. To expand the use of the MID unit concept, Johnston et al. later recommended using a distribution-based method to impute the MID when it was unknown [24]. First, they divided the MID by the sd of each study that used a scale with a known MID, to obtain an MID/sd ratio for each study. This provided a range of MID/sd ratios. They considered the median from this range of values as the "official" MID/sd ratio for the outcome measured by the scale. In the final step, for each study without an MID, they multiplied the sd by the official MID/sd ratio to obtain an estimate of the MID for that scale. This simple calculation provides an easy solution but includes significant assumptions. First, the distribution method assumes the MID/sd ratio is constant across different scales. Second, the uncertainty of the estimate from studies using a known MID is treated the same as the uncertainty of the estimate from studies using an imputed value for MID.

Our proposed solution is to consider the MID as a random variable instead of a constant. This approach addresses all of the limitations above. It enables investigators to obtain estimates of the mean difference standardized for MID (MD_{MID}) for questionnaires with no previous MID and to avoid making the unrealistic assumptions that 1) all people would provide the same value for MID, 2) the coefficient of variation for MID is independent of the measure, and 3) there is no correlation between the MID and the MD.

METHODS

Variance Calculations (individual studies)

The calculation for the variance of MD_{MID} is simply the variance of a ratio (MD / MID). When the MID is considered a constant, the variance of the ratio is simply:

$$\frac{variance_{MD}}{MID^2} \qquad \text{[eq.1]}$$

where $variance_{MD}$ is the variance of the difference in the group means, and MID is the proposed minimal important difference for the scale being used.

When the MID is considered to have a distribution of values (random variable), the distribution can be summarized with a mean ($mean_{MID}$) and sd (sd_{MID}). The variance of the ratio that represents MD_{MID} units (MD/MID) must now account for the correlation of the numerator and denominator of the ratio (rho), and the coefficient of variation (CoV) of the MID (sd_{MID} / $mean_{MID}$). Using the delta method and rearranging the equation to show the effect of CoV_{MID} (see Appendix 1 for derivation), the variance of the ratio is [30]:

$$Variance\left(\frac{MD}{MID}\right) =$$

$$\frac{\text{variance}_{\text{MD}}}{\text{mean}_{\text{MID}}{}^2} - \left[2 * \text{rho} * \text{sd}_{\text{MD}} * \text{CoV}_{MID} * \left(\frac{\text{MD}}{\text{mean}_{\text{MID}}{}^2}\right)\right] + \left[\left(\frac{\text{MD}}{\text{mean}_{\text{MID}}}\right)^2 * (\text{CoV}_{MID})^2\right] \quad \text{[eq.2]}$$

where $mean_{MID}$ is the mean of the MID distribution (equal to the value of MID when MID is considered a constant), rho is the correlation between $variance_{MD}$ and $mean_{MID}$, sd_{MD} is the pooled sd (between the two groups), and MD is the difference in group means.

In equation 2, when MID is a constant without a distribution, CoV_{MID} equals 0, the second and third terms each equal 0, and the formula collapses into equation 1. However, when MID has a distribution, equation 1 will underestimate or overestimate the variance depending on the balance of the second term (decreases variance) and third term (increases the variance) of equation 2. In brief, the variance will decrease as rho increases when CoV_{MID} is held constant, and the variance will increase as the CoV_{MID} increases when rho is held constant. Finally, everything else held constant, the effect on the variance increases as the mean difference between groups increases. This latter effect means greater care must be exercised when combining studies in a meta-analysis. Studies that use measures with poor sensitivity to change essentially have more measurement error or noise, which leads to higher sd (higher CoV_{MID} and higher variance) and a lower mean difference (lower variance). The overall effect on variance would depend on the balance of factors.

Empiric Data

Quality of Life in Chronic Obstructive Pulmonary Disease (COPD)

We replicated the data set used to first report a meta-analysis in MID units (Figure 1) [23]. The MID for the Chronic Respiratory Disease Questionnaire (CRQ) was reported as 0.5 on a 7-point scale, and the MID for the St. Georges Respiratory Questionnaire (SGRC) was reported as 4 on a 100-point scale.

Pain and Function in Patients with Osteoarthritis

We also analyzed data from a meta-analysis from Fransen et al. [25] This systematic review investigated the effect of exercise on pain and disability for participants with osteoarthritis in the knee. As seen in Appendix 2 and 3, the scale (and the range of the score) differs largely between the included studies, even within the same instrument measure.

RESULTS

We now present three illustrative meta-analysis examples. First, we present the health related quality of life following rehabilitation in COPD data that was used in the original description of the MID effect estimate (Figure 2 in Johnston et al.[23]), and show the results of sensitivity analyses varying rho and CoV_{MID}. Next, we apply the methods using data from meta-analyses investigating pain and disability following therapeutic exercise in knee osteoarthritis.

To calculate the variance, one must enter values for two variables that are not completely known and must be evaluated carefully using sensitivity analyses. First, we held rho constant between 0 and 1.0 in increments of 0.1, and varied CoV_{MID} between 0, 0.2, 0.5, 0.8 and 1. As an illustrative example, Figure 2 compares the square root of the variance (sd, in order to provide more spread between the points) between the two methods when rho is held constant at 0.5, and the CoV_{MID} equals 0.2, 0.5 and 0.8. Each study is represented by its corresponding letter from Figure 1. The line of identify shows where the points would fall if MID were a constant (if $CoV_{MID} = 0$, the two variance calculations are identical). The left

panel shows that the points are generally shifted to the left of the line of identity when CoV_{MID} = 0.2. This means the calculated variance when MID is considered a constant (line of identity) overestimates the true variance at CoV_{MID} = 0.2. This occurs because the second term in equation 2 will generally be greater than the third term of equation 2 if rho is large and CoV_{MID} is small. Examining the change in the position of the points as CoV_{MID} is increased, one sees the points generally move to the right, and the calculated variance when MID is considered a constant underestimates the true variance for the vast majority of studies. The effect is greatest when the treatment effect is greatest ("d" for SGRQ studies and "f", "g", "j" and "k" for the CRQ studies).

Second, we conducted sensitivity analyses by holding CoV_{MID} constant at 0.5 and calculating the variance when rho equaled 0, 0.5 and 0.8 (Figure 3). When rho is 0, the second term in equation 2 is 0, and the calculated variance when MID is considered a constant underestimates the true variance (points below the line of identity). As rho increases, the value of the second term in equation 2 increases and the points shift to the left.

In these analyses, we assumed the same CoV_{MID} for both measures for pedagogical reasons. If one measure had a greater sensitivity to change, the CoV_{MID} might be different between instruments. This could have important effects on the overall summary effect estimate. Our results are based on extraction of summary data from published meta-analyses and are only intended to illustrate the importance of considering MID as a distribution. They should not be interpreted as true estimates for the effects of the treatments proposed, which would require a complete systematic review, evaluation of the original papers, and characterization (or determination of what are realistic assumptions) of the distribution of MID for each measurement instrument.

Quality of Life in patients COPD, Pain and Disability in patients with Osteoarthritis

Table 1 presents the meta-analysis results for the Johnston et al. data [23], and the data examining the effects of exercise on osteoarthritis for pain, and for disability [11], when rho equals 0.5 and MID is varied between 0.2, 0.5 and 0.8. For the osteoarthritis data, we estimated the MID for each scale based on our understanding of the literature (Appendix 2 and 3). For each data set, we report the meta-analytical

summary effect estimates and 95% confidence intervals for fixed and random effects models, as well as heterogeneity statistics. For the COPD data, heterogeneity was small in all analyses. However, considering MID to have a distribution of values (with rho=0.5 and CoV_{MID} set to 0.8) instead of being a constant considerably 1) decreased the point estimate towards the null and 2) increased the 95% confidence intervals in both fixed and random effects models. In data from patients with osteoarthritis, the heterogeneity of the data was considered much greater when MID was considered a constant. Considering MID as a distribution instead of a constant also shifted the point estimates towards the null in both fixed and random effects models. Although the absolute width of the 95% confidence intervals in fixed effects models was considerably increased, it slightly decreased in random effects models.

Table 2 presents similar results to Table 1, and shows the changes in effect estimates and 95% confidence intervals when CoV_{MID} is held constant at 0.5, and rho is varied between 0.0, 0.5 and 0.8. The results and interpretations are similar to Table 1.

Presentation for Decision-Makers

In a recent GRADE article [26], the authors outlined different presentation options for continuous outcomes in summary of findings tables, including the use of MD_{MID} units. However, the choice for MD_{MID} presented listed only the comparative option (as the difference in MD_{MID}.) In fact, all presentation formats for SMD can be applied to MD_{MID} units, with similar but more transparent assumptions. In both cases, the outcome is a continuous scale with no limits on either end of the scale, and the approach is almost identical. We reproduced the results of Table 5 in Guyatt et al [26] as shown in Table 3, and added how one would provide similar results in MD_{MID} units (assumptions and calculations provided in Appendix 4), as well as some additional options we believe may be more meaningful to some patients. With identical presentation options, and the reduced risk of bias in estimating the treatment effect with MD_{MID} when MID is a random variable with a distribution, we believe the MD_{MID} approach should become the preferred method compared to the SMD approach when different continuous measures are used across studies to evaluate the same construct.

DISCUSSION

We have reviewed the meta-analytical implications that occur when the MID represents a distribution of values rather than a single value. Our approach avoids the assumptions that 1) MID should be considered constant across a single scale and for different patients, 2) the CoV_{MID} is independent of the measure and 3) there is no correlation between the MID and the MD of different scales. In sensitivity analyses using the original data describing MD_{MID} and setting the true rho at 0.5 and the true CoV_{MID} at 0.8, assuming MD_{MID} to be a constant would overestimate the treatment effect by 33% for COPD quality of life, and decrease the p-value for heterogeneity from 0.99 to 0.08. Similar or greater magnitudes of results were obtained in sensitivity analyses if rho=0.8 and CoV_{MID} =0.5 in this data set, and for studies examining pain and disability in patients with osteoarthritis. However, we highlight that CoV_{MID} were not available for the particular measures and we used the same value for each measure. Different results would be obtained if the CoV_{MID} were known and allowed to differ across measures, and there are minimal effects if rho and CoV_{MID} are small. This highlights the need for authors to comment on how realistic their assumptions are when imputing values for these variables, such as assuming MID is constant.

Meta-analyses of continuous outcomes represent a challenge because different investigators often use different scales to measure the same construct. For example, the meta-analysis for patients with osteoarthritis included 18 different scales for pain, and 12 different scales for disability. Although the SMD allows investigators to combine the results mathematically, there are important limitations as we described earlier. Perhaps more importantly, the ultimate objective of a meta-analysis is to help patients, clinicians and policy makers take decisions. The clinical interpretation of the SMD is difficult. What does a difference of 0.3 standard deviations really mean for a patient trying to evaluate if a treatment benefit is worth the associated side effects? Explaining that this is a small or moderate effect is not that helpful, even if one ignores the limitations of the arbitrary effect size cut-offs that are often applied [31]. The RoM approach proposed by Friedrich et al. [21, 22] is helpful if one believes the treatment acts on a

multiplicative scale. Explaining that pain is reduced by 50% is easily understandable, and is appropriate as long as the reduction is indeed 50% regardless of the baseline pain within the ranges studied. The MD_{MID} provides for a similar ease of interpretation and also incorporates a clinically meaningful perspective on magnitude of effect. Like the SMD, this is on the additive scale (treatment reduces pain by 2 points regardless of baseline value in the range of the studies) but without the challenges of the SMD.

In the original description on how to obtain an MID, the investigators questioned many participants and then used the mean value [32]. One way to think about the magnitude of potential bias is to think about signal-to-noise ratios. If MID is considered a constant, obtaining a precise estimate of the true MID is essential because even a small difference is recognized as bias. However, if there is some noise (i.e. variability/variance) in the signal, then small changes in the point estimate are somewhat masked by the noise/variance. Another perspective is that the variance in the MID measure is a combination of systematic error (bias), sampling error and inter-participant heterogeneity. With proper sampling techniques, there should not be systematic bias. The traditional method for estimating MID requires many participants to minimize sampling error. However, in reality, the total variance will usually be dominated by the inter-participant heterogeneity, and therefore increasing sampling size above a certain threshold (which would depend on context) will not have much of an effect. Therefore, optimal sample sizes of participants for eliciting MID will usually be considerably smaller, leading to easier methods and MID calculations for more instruments. More recently, Copay et al. reviewed several different methods to obtain an MID and all of them include simplifying a distribution into a single value. We believe acknowledging and incorporating the true variation in the MID is important because the variation partly reflects personal values and context, and partly reflects that MID could vary according to baseline values [28, 29], whether expressed on absolute or relative scale. Ignoring this variation and considering the MID as a constant equal to the mean of this distribution introduces significant challenges for calculating the variance of MD_{MID} in meta-analyses. First, it inappropriately suggests a precision that is not correct. Second, it requires omitting studies that do not have an MID already calculated, or using a method that assumes the MID/sd ratio is constant across all scales measuring the same construct [24].

Considering the MID as a random variable allows investigators to make more realistic assumptions about the distribution based on their substantive knowledge and clinical experience, and apply appropriate sensitivity analyses based on the mean, CoV_{MID} and rho.

We used the delta method to approximate the variance for the MD_{MID}. The delta method is commonly used to obtain an approximate variance for non-linear functions of random variables such as ratios [34, 35]. The focus of this paper was not a statistical assessment of optimality properties of estimators for parameters of interest, as one would typically do in methods performance assessment in the context of estimation as well as hypothesis testing. A comprehensive simulation study to assess optimality properties for estimators (e.g., bias and mean square error) and corresponding test statistics (e.g., Type I error and statistical power) was beyond the scope of this paper. However, there are some known limitations when ratios are the functions of interest, as in our application. First, complications can arise when the denominator gets close to zero. Second, statistical inference (e.g., confidence intervals) assumes that the bivariate random vector consisting of the numerator and denominator is distributed as bivariate normal (exactly or approximately). This assumption is likely to be violated when sample size is small or when there are outliers. Third, the effect we observed due to CoV_{MID} and rho is greatest for those studies with the largest mean difference. Studies using measures that are sensitive to change would be expected to have higher mean differences, which might lead to a greater variance and inappropriately down weighting. For example, increasing CoV_{MID} had the greatest effect on the variance of the Behnke 2000 study (study "f"), which means this study (with a large effect) might receive a lower weight when MID is considered random variable. However, these studies with greater sensitivity to change are also expected to have a smaller CoV_{MID} because there is less "noise" or error in the measure of improvement, which leads to a reduced variance and up weighting. Which of these two effects would predominate, and the overall effect on the relative weighting of studies in any particular meta-analysis, is difficult to predict. Fourth, regardless of how one estimates the effect in MID units (e.g. anchor based methods, distribution based methods), one must always estimate a value for rho and CoV_{MID}, which may be from observed data, substantive knowledge or guessing. Although we applied the same CoV_{MID} and rho to each measure in

our sensitivity analyses to illustrate why it is important to consider the MID as a random variable with a distribution, we emphasize that best practice would be to estimate a separate CoV_{MID} and rho for each measure, through either observed data or substantive knowledge. If observed data or substantive knowledge are not available, then investigators should decide if it is reasonable to assume a value of 0 for both variables. The methods described in this paper can be used to estimate how the results may be under or overestimated, should the true values be different from 0. Finally, if sensitivity to change is quite different for different measures, investigators should carefully consider whether it is appropriate to combine these studies in a meta-analysis, whether MID is considered a constant or as having a distribution.

CONCLUSIONS

The objective of a meta-analysis is to help make informed decisions. The MD_{MID} is an easily interpretable measure of effect for patients, clinicians and decision makers. Considering the MID as a random variable with a mean and sd instead of a constant allows for a more appropriate estimate of the variance, and thus a more appropriate estimate of treatment effects.

ABBREVIATIONS

EBM: Evidence-based medicine

WOMAC: Western Ontario and McMaster University Osteoarthritis Index

KOOS: Knee Injury and Osteoarthritis Outcome Score

HOOS: Hip Disability and Osteoarthritis Outcome Score

HAQ: Health Assessment Questionnaire

AIMS: Arthritis Impact Measurement Scales

MD: Mean Difference (Mean Group 1 – Mean Group 2)

sd: standard deviation

SMD: Standardized Mean Difference (MD/sd) is the mean difference expressed in sd units

RoM: Ratio of Means (Mean Group 1 / Mean Group 2)

MID: Minimal Important Difference

MD_{MID}: Mean Difference expressed in MID units

CoV_{MID}: coefficient of variation of the MID

sd_{MID}: standard deviation of the MID distribution

$mean_{MID}$: mean of the MID distribution

$sd_{MeanDiff}$: pooled sd of the two groups being compared

MeanDiff: mean difference

COPD: Chronic obstructive pulmonary disease

CRQ: Chronic respiratory disease questionnaire

SGRC: St. Georges Respiratory questionnaire

ACKNOWLEDGEMENTS

This study received no specific funding. IS was supported by the Lady Davis Institute, Jewish General Hospital in Montreal Canada. RC (The Parker Institute) was supported by unrestricted grants from the Oak Foundation; CJ was supported by the University of Southern Denmark and the University Hospital of Copenhagen, Herlev and Gentofte. JB is funded by the Canadian Institutes of Health Research. None of the funders had any involvement in the conduct of this study.

REFERENCES

1. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. Br Med J 1996;312(7023):71-72.
2. Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. Br Med J 1995;310(6987):1122-1126.
3. Mountokalakis TD. Evidence-based medicine vs inferential reasoning: the case of hypertension associated with renal disease. Nephrol Dial Transplant 2001;16 Suppl 6:4-6.
4. Clarke M, Langhorne P. Revisiting the Cochrane Collaboration. Meeting the challenge of Archie Cochrane--and facing up to some new ones. Br Med J 2001;323(7317):821.
5. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Stat Med 1999;18(3):321-359.
6. Egger M, Smith GD, O'Rourke K. Rationale, potentials, and promise of systematic reviews. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in health care Meta-analysis in context. London: BMJ Publishing Group, 2001:3-19.
7. Guyatt GH, Oxman AD, Santesso N, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66(2):158-172.
8. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. Br Med J 2010;340:c869.
9. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21(11):1575-1600.
10. Ghogomu EA, Maxwell LJ, Buchbinder R, et al. Updated method guidelines for cochrane musculoskeletal group systematic reviews and metaanalyses. J Rheumatol 2014;41(2):194-205.

11. Juhl C, Lund H, Roos EM, et al. A hierarchy of patient-reported outcomes for meta-analysis of knee osteoarthritis trials: empirical evidence from a survey of high impact journals. Arthritis 2012;2012:136245.

12. Bellamy N, Buchanan WW, Goldsmith CH, et al. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. J Rheumatol 1988;15(12):1833-1840.

13. Roos EM, Roos HP, Lohmander LS, et al. Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. J Orthop Sports Phys Ther 1998;28(2):88-96.

14. Nilsdotter AK, Lohmander LS, Klassbo M, et al. Hip disability and osteoarthritis outcome score (HOOS)--validity and responsiveness in total hip replacement. BMC Musculoskelet Disord 2003;4:10.

15. Fries JF, Spitz P, Kraines RG, et al. Measurement of patient outcome in arthritis. Arthritis Rheum 1980;23(2):137-145.

16. Lequesne M. Indices of severity and disease activity for osteoarthritis. Semin Arthritis Rheum 1991;20(6 Suppl 2):48-54.

17. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis. The arthritis impact measurement scales. Arthritis Rheum 1980;23(2):146-152.

18. Meenan RF, Mason JH, Anderson JJ, et al. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum 1992;35(1):1-10.

19. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain 1975;1(3):277-299.

20. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Available from http://www.cochrane-handbook.org, 2011.

21. Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. J Clin Epidemiol 2011;64(5):556-564.

22. Friedrich JO, Adhikari NK, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. BMC Med Res Methodol 2008;8:32.

23. Johnston BC, Thorlund K, Schunemann HJ, et al. Improving the interpretation of quality of life evidence in meta-analyses: the application of minimal important difference units. Health Qual Life Outcomes 2010;8:116.

24. Johnston BC, Thorlund K, da Costa BR, et al. New methods can extend the use of minimal important difference units in meta-analyses of continuous outcome measures. J Clin Epidemiol 2012;65(8):817-826.

25. Fransen M, McConnell S. Exercise for osteoarthritis of the knee. Cochrane Database Syst Rev 2008(4):CD004376.

26. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66(2):173-183.

27. van den Noortgate W, Onghena P. Estimating the mean effect size in meta-analysis: Bias, precision, and mean squared error of different weighting methods. Behav Res Meth Instr 2003;35(4):504-511.

28. Rouquette A, Blanchin M, Sebille V, et al. The minimal clinically important difference determined using item response theory models: an attempt to solve the issue of the association with baseline score. J Clin Epidemiol 2014;67(4):433-440.

29. Zhang Y, Zhang S, Thabane L, et al. Although not consistently superior, the absolute approach to framing the minimally important difference has advantages over the relative approach. J Clin Epidemiol 2015;68(8):888-894.

30. Knight K. Mathematical statistics. Unites States: Chapman & Hall/CRC, 1999.

31. Bliddal H, Christensen R. The treatment and prevention of knee osteoarthritis: a tool for clinical decision-making. Expert Opin Pharmacother 2009;10(11):1793-1804.

32. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials 1989;10(4):407-415.

33. Copay AG, Subach BR, Glassman SD, et al. Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 2007;7(5):541-546.

34. Properties of a Random Sample. In. Statistical Inference: Duxbury Thomson Learning, 2002:240-245.

35. Oehlert GW. A Note on the Delta Method. The American Statistician 1992;46(1):27-29.

36. Stovitz SD, Shrier I. Medical decision making and the importance of baseline risk. Br J Gen Pract 2013;63:795-797.

37. Furukawa TA. From effect size into number needed to treat. Lancet 1999;353(9165):1680.

38. Suissa S. Binary methods for continuous outcomes: a parametric alternative. J Clin Epidemiol 1991;44(3):241-248.

Figure Legends

Figure 1. Meta-analysis of health related quality of life for rehabilitation of chronic obstructive pulmonary disease based on data presented in Johnston et al. [23] with the effect estimate presented as a mean difference between groups using a scale of MID units. The questionnaires were the St. Georges Respiratory Questionnaire (SGRC: MID reported as 4 points on a 100-point scale) and Chronic Respiratory Disease Questionnaire (CRQ: MID reported as 0.5 on a 7-point scale). Our figure is minimally different from Figure 2 in Johnston et al because we used the raw numbers in their Figure 1 to generate the meta-analysis (there were small discrepancies between "n" in Figures 1 and 2), and also corrected some minimal mathematical errors.

Figure 2: The square root of the variance (sd, to provide more spread between the points) when MID (minimal important difference) is a constant is plotted against the square root of the variance (delta method) when MID is considered a random variable, when the correlation between the mean difference between groups and the MID is held constant at 0.5, and the coefficient of variation for MID (sd of MID / mean MID) is varied between 0.2, 0.5 and 0.8. The studies are labeled with letters corresponding to the numbers with each study in Figure 1. Points that are below the line of identity indicate that considering MID as a constant will underestimate (and points above the line of identity will overestimate) the variance given the values of rho and coefficient of variation for MID of the specified simulation.

Figure 3. The square root of the variance (sd, to provide more spread between the points) when MID (minimal important difference) is a constant is plotted against the square root of the variance (delta method) when MID is considered a random variable, when the correlation between the mean difference between groups and the MID (rho) is varied between 0, 0.5 and 0.8, and the coefficient of variation for MID (sd of MID / mean MID) is held constant at 0.5.The studies are labeled with letters corresponding to the numbers with each study in Figure 1. Points that are below the line of identity indicate that

considering MID as a constant will underestimate (and points above the line of identity will overestimate) the variance given the values of rho and coefficient of variation for MID of the specified simulation.

Table 1. The effect of varying the coefficient of variation of minimal important difference (CoV_{MID}) on meta-analysis results[#] when the correlation (rho) between the minimal important difference (MID) and mean difference is equal to 0.5. Outcome measures include changes in quality of life with treatment for chronic obstructive pulmonary disease (COPD-QOL) (data from Johnson et al [23]), changes in Pain (OA-Pain) and Disability (OA-Disability) with exercise for osteoarthritis (data from [25]).

					Heterogeneity Statistics		
Outcome	Rho	CoV_{MID}	Fixed effect	Random effect	I-squared	Tau-squared	p-value
COPD-QOL	MID as a Constant		1.79 (1.51 to 2.07)	1.74 (1.35 to 2.12)	35%	0.1895	0.08
	0.5	0.2	1.67 (1.40 to 1.94)	1.69 (1.34 to 2.03)	29%	0.1319	0.13
	0.5	0.5	1.47 (1.07 to 1.87)	1.47 (1.07 to 1.87)	0%	0	0.95
	0.5	0.8	1.35 (0.77 to 1.92)	1.35 (0.77 to 1.92)	0%	0	0.99
OA-Pain	MID as a Constant		-0.13 (-0.16 to -0.11)	-0.29 (-0.36 to -0.21)	77%	0.0245	<0.0001
	0.5	0.2	-0.11 (-0.15 to -0.08)	-0.21 (-0.28 to -0.14)	48.8%	0.0125	0.0012
	0.5	0.5	-0.09 (-0.14 to -0.04)	-0.09 (-0.14 to -0.04)	0%	0	0.56
	0.5	0.8	-0.08 (-0.15 to -0.02)	-0.08 (-0.15 to -0.02)	0%	0	0.98
OA-Disability	MID as a Constant		-0.21 (-0.25 to -0.17)	-0.28 (-0.37 to -0.18)	78.3%	0.0428	<0.0001
	0.5	0.2	-0.16 (-0.21 to -0.11)	-0.19 (-0.27 to -0.12)	36.4%	0.0115	0.024
	0.5	0.5	-0.12 (-0.19 to -0.06)	-0.12 (-0.19 to -0.06)	0%	0	0.84

	0.5	0.8	-0.10 (-0.17 to -0.02)	-0.10 (-0.17 to -0.02)	0%	0	0.99

[#] Results are expressed as mean differences in MID units with 95% confidence intervals for both fixed effects and random effects models, along with I-squared, tau-squared and the p-value for heterogeneity. When MID is considered a constant, rho and CoV equal 0, and therefore only one value is provided for each outcome. When rho was set to 0.8 (results not shown), the results were similar with only minimal decreases in the effect estimates and measures of heterogeneity, although the 95% confidence interval for rho=0.8 and ratio=0.8 reached or crossed 0 for osteoarthritis outcomes [fixed and random effect estimates -0.05 (-0.11 to 0.01) for pain; fixed and random effect estimates -0.08 (-0.15 to 0.00) for disability].

Table 2. The effect of varying the correlation (rho) between the minimal important difference (MID) and mean difference on meta-analysis results[#], when coefficient of variation of minimal important difference (CoV_{MID}) is equal to 0.5. Outcome measures include changes in quality of life with treatment for chronic obstructive pulmonary disease (COPD-QOL) (data from Johnson et al publication), changes in Pain (OA-Pain) and Disability (OA-Disability) with exercise for osteoarthritis.

					Heterogeneity Statistics		
Outcome	Rho	CoV_{MID}	Fixed effect	Random effect	I-squared	Tau-squared	p-value
COPD-QOL	MID as a constant		1.79 (1.51 to 2.07)	1.74 (1.35 to 2.12)	35%	0.1895	0.08
	0.2	0.5	1.61 (1.26 to 1.97)	1.61 (1.26 to 1.97)	0%	0	0.62
	0.5	0.5	1.43 (0.90 to 1.96)	1.43 (0.90 to 1.96)	0%	0	0.98
	0.8	0.5	1.28 (0.57 to 2.00)	1.28 (0.57 to 2.00)	0%	0	0.99
OA-Pain	MID as a constant		-0.13 (-0.16 to -0.11)	-0.29 (-0.36 to -0.21)	77%	0.0245	<0.0001
	0.2	0.5	-0.12 (-0.15 to -0.09)	-0.24 (-0.31 to -0.17)	61.6%	0.0158	<0.0001
	0.5	0.5	-0.10 (-0.15 to -0.06)	-0.15 (-0.21 to -0.09)	14.2%	0.0032	0.24
	0.8	0.5	-0.10 (-0.15 to -0.04)	-0.10 (-0.15 to -0.04)	0%	0	0.95
OA-Disability	MID as a constant		-0.21 (-0.25 to -0.17)	-0.28 (-0.37 to -0.18)	78.3%	0.0428	<0.0001
	0.2	0.5	-0.17 (-0.21 to -0.13)	-0.22 (-0.30 to -0.15)	51.8%	0.0165	0.0005
	0.5	0.5	-0.14 (-0.20 to -0.09)	-0.14 (-0.20 to -0.09)	0%	0	0.6048

	0.8	0.5	-0.12 (-0.19 to -0.05)	-0.12 (-0.19 to -0.05)	0%	0	0.98

[#] Results are expressed as mean differences in MID units with 95% confidence intervals for both fixed effects and random effects models, along with I-squared, tau-squared and the p-value for heterogeneity. When MID is considered a constant, rho and CoV equal 0, and therefore only one value is provided for each outcome.

Table 3: Different recommended formats for the GRADE Summary of Findings tables [26] using health related quality of life data from Johnston as the example. Italicized rows are new suggestions not represented in [26] based on the premise that decision-makers should be given two numbers to compare, rather than one number and a comparative estimate that has to be converted to a second number [36].

Outcome	Estimated baseline score/ proportion improving in control patients	Estimated baseline score/ proportion improving in treatment patients	Certainty of the Evidence	Comments
<u>Comparative Estimate</u>				
Quality of Life (sd units)	The HRQL score in the respiratory rehabilitation group improved on average 0.73 (95%CI: 0.49 to 0.96) more sd units in the respiratory rehabilitation patients than in the control patients		♁♁♁♁ High	As a rule of thumb, 0.2 sd represents a small difference, 0.5 a moderate, and 0.8 a large
Quality of Life (MID units)	The HRQL score in the respiratory rehabilitation group improved on average an extra 1.47 (1.07 to 1.87) MID units compared to control patients		♁♁♁♁ High	Estimated assuming a moderate (0.5) correlation between variation in MID and the mean difference on a scale, and moderate variability in what is considered the MID itself.
Mean of each Group				
Quality of Life (from sd units and HRQOL scale 1-7)[a]	Control group baseline, 4.5[a] Average improvement in control was 0.04	HRQL improved on average 0.73 (95%CI: 0.49 to 0.96) sd units more in the respiratory rehabilitation patients than in the control patients	♁♁♁♁ High	Calculated by transforming all scores to the CRQ in which the minimally important difference is 0.5, and multiplying 0.01 sd unit mean by 7
Quality of Life (MID units)[a]	Average improvement in control was 0.08 MID units	HRQL improved on average 1.47 (95%CI: 1.07 to 1.87) MID units more in the respiratory rehabilitation patients than in the control patients	♁♁♁♁ High	Average in control group calculated as the median of all control group scores expressed in MID units
Quality of Life (MID units)[a]	*Average improvement in control was 0.08 MID units*	*Average improvement in treated group was 1.55 (95%CI: 1.15 to 1.95) MID units[b]*	*♁♁♁♁ High*	*Average in control group calculated as the median of all control group scores expressed in MID units*
Proportion of Patients in each Group with Important Improvement [c]				

Based on sd units	0.3[d]	Differences in proportion achieving important improvement 0.31 (95% CI: 0.22, 0.40) in favor of rehabilitation	⚷⚷⚷⚷ High	Calculation uses established minimally important difference of 0.5 units on the CRQ and 4 units on the St. George's Respiratory Questionnaire
Based on MID units	0.18 *achieved important improvement*	Differences in proportion achieving important improvement 0.53 (95%CI: 0.48 to 0.65) in favor of rehabilitation [e]	⚷⚷⚷⚷ High	Calculation uses 1 MID as the definition for Important Improvement and assumes a normal distribution for subjects' scores.
Quality of Life (MID units)	*0.18 achieved important improvement*	*0.71 (95%CI: 0.56 to 0.83) achieved important improvement* [b]	⚷⚷⚷⚷ *High*	*Calculation uses 1 MID as the definition for Important Improvement and assumes a normal distribution for subjects' scores.*

[a] Approximate average of baseline control group scores in the studies that reported the baseline score (numbers from [26]).

[b] Calculated by summing the control group improvement and the treatment effect

[c] The assumptions underlying these calculations were not provided in [26]. See Appendix 4 for details.

[d] This represents the median of the proportion of patients in the control group who achieved an important improvement. That is, the proportion with improvement was calculated for each study [more than 0.5 (CRQ) or 4 (St. George's)], and the median of these values was 0.3, suggesting 30% of the control group achieved an important improvement.

Appendix 1: Using the delta method, the variance of the ratio is traditionally written as [30]

$$Variance\left(\frac{MD}{MID}\right) = \frac{\text{variance}_{\text{MD}}}{\text{mean}_{\text{MID}}{}^{2}} - \left[2 * \text{rho} * \text{sd}_{\text{MD}} * \text{sd}_{\text{MID}} * \left(\frac{\text{MD}}{\text{mean}_{\text{MID}}{}^{3}}\right)\right] + \frac{\text{MD}^{2} * \text{sd}_{\text{MID}}{}^{2}}{\text{mean}_{\text{MID}}{}^{4}} \quad \text{[eq.A1]}$$

where MD is the mean difference between groups, MID is the proposed minimal important difference for the scale being used, $\text{variance}_{\text{MD}}$ is the variance of the difference in the group means, mean_{MID} is the mean of the MID distribution (equal to the value of MID when MID is considered a constant), rho is the correlation between $\text{variance}_{\text{MD}}$ and mean_{MID}, and sd_{MD} is the pooled sd (between the two groups). Rearranging the formula, highlights the effect of CoV_{MID}:

$$\frac{\text{variance}_{\text{MD}}}{\text{mean}_{\text{MID}}{}^{2}} - \left[2 * \text{rho} * \text{sd}_{\text{MD}} * \frac{\text{sd}_{\text{MID}}}{\text{mean}_{\text{MID}}} * \left(\frac{\text{MD}}{\text{mean}_{\text{MID}}{}^{2}}\right)\right] + \left[\left(\frac{\text{MD}}{\text{mean}_{\text{MID}}}\right)^{2} * \left(\frac{\text{sd}_{\text{MID}}}{\text{mean}_{\text{MID}}}\right)^{2}\right] \quad \text{[eq.A2]}$$

$$\frac{\text{variance}_{\text{MD}}}{\text{mean}_{\text{MID}}{}^{2}} - \left[2 * \text{rho} * \text{sd}_{\text{MD}} * \text{CoV}_{MID} * \left(\frac{\text{MD}}{\text{mean}_{\text{MID}}{}^{2}}\right)\right] + \left[\left(\frac{\text{MD}}{\text{mean}_{\text{MID}}}\right)^{2} * (\text{CoV}_{MID})^{2}\right] \quad \text{[eq.A3]}$$

Appendix 2: Values of minimal important difference (MID) for studies on examining the effect of exercise on pain in patients with osteoarthritis.

Author	Year	Scale	MID
Minor	1989	AIMS (sum of 4 items standardized; range 0-10)	2
Kovar	1992	AIMS (1 ordinal scale standardized; range 0-10)	3
Ettinger (a)	1997	CPS (mean of 6 items; range 1-6)	1.5
Ettinger (b)	1997	CPS (mean of 6 items; range 1-6)	1.5
Bautch	1997	VAS (mean of 2 VAS; range 0-10)	2
Rogind	1998	NRS (one 11 box pain scale; range 0-10)	3
van Baar	1998	VAS (1 VAS; range 0-100)	20
Peloquin	1999	AIMS 2 (mean of 5 items standardized; range 0-10)	2
Maurer	1999	WOMAC (sum of 5 VAS; range 0-500)	100
O'Reilly	1999	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Hopman-Rock	2000	VAS (1 VAS; range 0-100)	20
Deyle	2000	WOMAC (sum of 5 VAS; range 0-500)	100
Petrella	2000	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Baker	2001	WOMAC (sum of 5 VAS; range 0-500)	100
Fransen	2001	WOMAC (mean of 5 items scored 0 to 4; range 0-100)	20
Gur	2002	NRS11 (sum of 7 11-box NRS scored 0-10; range 0-70)	21
Thomas	2002	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Topp	2002	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Talbot	2003	McGill Pain Questionnaire (1 ordinal scale; range 0-5)	1
Huang	2003	VAS (Walk-Stand) (1 VAS; range 0-10)	2
Quilty	2003	VAS (Pain Overall) (1 item scored 0-100)	20

Foley	2003	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Song	2003	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Keefe	2004	AIMS (mean of 4 items standardized; range 0-10)	2
Hughes	2004	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Messier	2004	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Thorstensson	2005	KOOS (9 items scored 0-4, standardized; range 0-100)	10
Huang a	2005	VAS (Walk-Stand) (1 item scored 0-10)	2
Bennell	2005	VAS (Pain Move) (1 item scored 0-10)	2
Hay	2006	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Mikesky	2006	WOMAC (sum of 5 items scored 0 to 4; range 0-20)	6
Fransen	2007	WOMAC (mean of 5 items scored 0 to 4; range 0-100)	30

Appendix 3: Values of minimal important difference (MID) for studies on examining the effect of exercise on disability in patients with osteoarthritis.

Author	Year	Scale	MID
Minor	1989	AIMS (physical activity) (sum of 5 items standardized; range 0-10)	2
Kovar	1992	AIMS (physical activity) (sum of 5 items standardized; range 0-10)	2
Schilke	1996	OASI (mobility) (mean of VAS; range 0-10)	1.7
Bautch	1997	AIMS (disability) (sum of 45 items scored 0-1; range 0-45)	13.5
Ettinger a	1997	23 questions on disability (mean of 23 items; range 1-5)	1.5
Ettinger b	1997	23 questions on disability (mean of 23 items; range 1-5)	1.5
Rogind	1998	Algofunctional Index (AFI) (sum of 10 items; range 0-24)	3.5
van Baar	1998	Impact of rheumatic diseases on health and lifestyle (sum of 7 items scored 1-4; range 7-28)	4.4
Maurer	1999	WOMAC Disability (sum of 17 items scored 0-100; range 0-1700)	340
O'Reilly	1999	WOMAC Disability (sum of 7 items scored 1-4; range 7-28)	13.6
Peloquin	1999	AIMS 2(walking and bending) (mean of 5 items standardized; range 0-10)	2
Deyle	2000	WOMAC Disability (sum of 17 items scored ; range 0-1700)	340
Hopman-Rock	2000	Impact of rheumatic diseases on health and lifestyle (mobility) (sum of 7 items scored 1-4; range 7-28)	4.4
Petrella	2000	WOMAC Disability (100 mm scale) (mean of 17 VAS; range 0-10)	2
Baker	2001	WOMAC Disability (sum of 17 VAS; range 0-1700)	340
Fransen	2001	WOMAC Disability (mean of 17 VAS; range 0-100)	20
Gur	2002	NRS (sum of 7 items scored 0-10; range 0-70)	15
Thomas	2002	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Topp	2002	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6

Foley	2003	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Huang	2003	Lequesne (sum of 11 items; range 0-26)	5.2
Quilty	2003	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Song	2003	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Hughes	2004	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Messier	2004	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Bennell	2005	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Huang a	2005	Lequesne (sum of 11 items; range 0-26)	5.2
Thorstensson	2005	KOOS (ADL) (sum of 17 items standardized; range 0-100)	20
Hay	2006	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Mikesky	2006	WOMAC Disability (sum of 17 items scored 0-4; range 0-68)	13.6
Fransen	2007	WOMAC Disability (mean of 17 items scored; range 0-100))	20

Appendix 4: Explanation of assumptions for presentation of proportions in the GRADE Summary of Findings table.

One method to calculate the proportion improved for a control group in a single study is:

$$\Phi\left[\frac{(C - mean_{Control})}{sd_{Control}}\right]$$

where Φ is the cumulative standard normal distribution, C is the cutoff threshold, and $mean_{Control}$ and $sd_{Control}$ are the mean and sd of the control group respectively [37, 38]. In brief, this equation is based on a standard normal distribution with mean equal to 0 and sd equal to 1. The cutoff threshold is then shifted to what it would have been for this distribution by subtracting the mean of the observed data from the threshold. One then calculates the probability of being above the adjusted threshold. This is equivalent to creating a cumulative normal distribution with mean and sd equal to the observed data, and calculating the probability of being above the threshold.

We now have to expand this concept to accommodate meta-analysis. The first step is to estimate the overall control mean and sd for all the studies combined. From the footnote in Table 5 in the GRADE paper [26], it appears the authors calculated the probability of improvement for each study, and then took the median of these probabilities. However, this approach gives equal weight to each study and ignores the normal meta-analytical approach of weighting studies. Alternatively, one could take a weighted mean or weighted median of the results in the controls and use these values for the calculation of probability of improvement. Because our purpose is simply to demonstrate the presentation of results and rather than reporting valid estimates, we used the apparent methods in [26] in order to avoid confusion.

We now turn our attention to the proportion improved in the treatment group. Simply applying the same methods to the treatment group as one did for the control group would effectively be ignoring randomization and treating the control and treatment groups as completely separate studies. The methods

in [26] are not provided for calculating proportion improved when using MID units. We were able to approximate their results (0.32 increased improvement versus 0.31 increased improvement) by considering the treatment mean to be the difference in means (0.71) plus the control mean (0.04). However, this method does not require the comment in the table detailing the MID for CRQ and SGRC questionnaires.

For the calculations in our table using only MID units, we first calculated the mean change in MID units for the control group using the same unweighted methods in [26] in order to be consistent (0.08 MD_{MID} units). We then calculated the proportion improved in the control group considering MID equal to 1 as the important difference and an sd equal to 1. The mean of the treatment group is calculated by adding the treatment effect in MD_{MID} units with rho = 0.5 and CoV = 0.5 [1.47 (1.07 to 1.87)] to the mean of the control group (0.08), which yields was 1.55 (95%CI: 1.15 to 1.95). We then calculated the overall proportion improved in the treatment group using these mean values, and the overall difference in proportions improved was simply the difference between the proportion improved in the treatment group and the proportion improved in the control group.